UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/813,642	03/30/2004	Ara V. Nefian	884.C05US1	4943

21186 7590 06/04/2007

SCHWEGMAN, LUNDBERG, WOESSNER & KLUTH, P.A.
P.O. BOX 2938
MINNEAPOLIS, MN 55402

EXAMINER	
STOFFREGEN, JOEL	

ART UNIT	PAPER NUMBER
2626	

MAIL DATE	DELIVERY MODE
06/04/2007	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	10/813,642	NEFIAN ET AL.	
	Examiner	Art Unit	
	Joel Stoffregen	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *30 March 2004*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-28* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-28* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *30 March 2004* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date *08/11/2005 and 05/24/2006*.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

1. This action is in response to the original application filed on 03/30/2004.

2. Claims 1-28 are currently pending in this application. Claims 1, 8, 15, 20, and 25 are independent claims.

Information Disclosure Statement

3. The information disclosure statements (IDS) submitted on 08/11/2005 and 05/24/2006 are being considered by the examiner.

Claim Rejections - 35 USC § 102

4. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of application for patent in the United States.

5. **Claims 1, 2, 8, 11, 13, 15, 16, 20-22, 24, 25, and 28** are rejected under 35 U.S.C. 102(b) as being anticipated by Katsumi Patent No.: US 6,369,846 ("KATSUMI").

6. Regarding **claim 1**, KATSUMI teaches a method, comprising:

electronically capturing visual features associated with a speaker speaking ("speaking attendee determination information based on video signal", KATSUMI, column 6, lines 52-53);

electronically capturing audio ("speaking attendee determination information based on audio signal", KATSUMI, column 6, lines 50-51);

matching selective portions of the audio with the visual features ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63); and

identifying the remaining portions of the audio as potential noise not associated with the speaker speaking ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

7. Regarding **claim 2**, KATSUMI further teaches:

electronically capturing additional visual features associated with a different speaker speaking ("images of terminals determined as having speaking attendees", KATSUMI, column 7, lines 56-57); and

matching some of the remaining portions of the audio from the potential noise with the additional speaker speaking ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

8. Regarding **claim 8**, KATSUMI teaches a method, comprising:

monitoring an electronic video of a first speaker and a second speaker ("images of terminals determined as having speaking attendees", KATSUMI, column 7, lines 56-57);

concurrently capturing audio associated with the first and second speaker speaking ("voices may be contained in the audio signal", KATSUMI, column 3, line 1);

analyzing the video to detect when the first and second speakers are moving their respective mouths ("extracts the change amount of the shape of the lip portion", KATSUMI, column 6, lines 24-25); and

matching portions of the captured audio to the first speaker and other portions to the second speaker based on the analysis ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

9. Regarding **claim 11**, KATSUMI further teaches separating the electronic video from the concurrently captured audio in preparation for analyzing (see KATSUMI, FIG. 3, the audio processing is separate from the video processing).

10. Regarding **claim 13**, KATSUMI further teaches identifying selective portions of the captured audio as noise if the selective portions have not been matched to the first speaker or the second speaker ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

11. Regarding **claim 15**, KATSUMI teaches a system, comprising:

a camera (see KATSUMI, column 4, lines 22-23, the conference terminals produce video signals, therefore a camera is inherent);

a microphone (see KATSUMI, column 4, lines 22-23, the conference terminals produce audio signals, therefore a microphone is inherent); and

a processing device ("MCU", KATSUMI, column 4, line 21), wherein the camera captures video of a speaker and communicates the video to the processing device, the microphone captures audio associated with the speaker and an environment of the speaker and communicates the audio to the processing device ("the conference

terminals 6a to 6c multiplex video signals and audio signals of locations [A] to [C]... and transmit the transmission signals to the MCU", KATSUMI, column 4, lines 22-26), the processing device includes instructions that identifies visual features of the video where the speaker is speaking ("speaking attendee determination information based on video signal", KATSUMI, column 6, lines 52-53) and uses time dependencies to match portions of the audio to those visual features ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63).

12. Regarding **claim 16**, KATSUMI further teaches that the captured video also includes images of a second speaker ("images of terminals determined as having speaking attendees", KATSUMI, column 7, lines 56-57) and the audio includes sounds associated with the second speaker ("voices may be contained in the audio signal", KATSUMI, column 3, line 1), and wherein the instructions matches some portions of the audio to the second speaker when some of the visual features indicate the second speaker is speaking ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

13. Regarding **claim 20**, KATSUMI teaches a machine accessible medium having associated instructions, which when accessed, results in a machine performing:

separating audio and video associated with a speaker speaking (see KATSUMI, FIG. 3, the audio processing is separate from the video processing);

identifying visual features from the video that indicate a mouth of the speaker is moving or not moving ("extracts the change amount of the shape of the lip portion", KATSUMI, column 6, lines 24-25); and

associating portions of the audio with selective ones of the visual features that indicate the mouth is moving ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63).

14. Regarding **claim 21**, KATSUMI further teaches including instructions for associating other portions of the audio with different ones of the visual features that indicate the mouth is not moving ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

15. Regarding **claim 22**, KATSUMI further teaches instructions for:

identifying second visual features from the video that indicate a different mouth of another speaker is moving or not moving ("images of terminals determined as having speaking attendees", KATSUMI, column 7, lines 56-57); and

associating different portions of the audio with selective ones of the second visual features that indicate the different mouth is moving ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

16. Regarding **claim 24**, KATSUMI further teaches that the instructions for associating further include instructions for matching same time slices associated with a time that the portions of the audio were captured and the same time during which the selective ones of the visual features were captured within the video ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63).

17. Regarding **claim 25**, KATSUMI teaches an apparatus, residing in a computer-accessible medium, comprising:

face detection logic ("detects at least the lip portion of a conference attendee from the video signal", KATSUMI, column 6, lines 23-34);

mouth detection logic ("extracts the change amount of the shape of the lip portion", KATSUMI, column 6, lines 24-25); and

audio-video matching logic, wherein the face detection logic detects a face of a speaker within a video ("detects at least the lip portion of a conference attendee", KATSUMI, column 6, lines 23-34), the mouth detection logic detects and monitors movement and non-movement of a mouth included within the face of the video ("extracts the change amount of the shape of the lip portion", KATSUMI, column 6, lines 24-25), and the audio-video matching logic matches portions of captured audio with any movements identified by the mouth detection logic ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63).

18. Regarding **claim 28**, KATSUMI further teaches that the apparatus resides on a processing device and the processing device is interfaced to a camera and a microphone (see KATSUMI, column 4, lines 22-23, the conference terminals produce audio and video signals, therefore a camera and microphone are inherent).

Claim Rejections - 35 USC § 103

19. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

20. **Claims 3-6, 9, 10, 14, 17-19, 23, 26, and 27** are rejected under 35 U.S.C. 103(a) as being unpatentable over Katsumi Patent No.: US 6,369,846 ("KATSUMI") in view of Nefian Pub. No.: US 2003/0212557 ("NEFIAN").

21. Regarding **claim 3**, KATSUMI teaches all of the limitations of claim 1.

However, KATSUMI does not disclose generating parameters associated with the matching and the identifying and providing the parameters to a Bayesian Network which models the speaker speaking.

In the same field of audiovisual processing, NEFIAN teaches generating parameters associated with the matching and the identifying ("audio processing and visual feature extraction", NEFIAN, paragraph [0012]) and providing the parameters to a Bayesian Network which models the speaker speaking ("video data must be fused with audio data using... a coupled hidden Markov model [HMM]", NEFIAN, paragraph [0023], where the HMM is a dynamic Bayesian network).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual matching method of NEFIAN

with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003])

22. Regarding **claim 4**, NEFIAN further teaches that electronically capturing the visual features further includes processing a neural network ("neural network", NEFIAN, paragraph [0014]) against electronic video associated with the speaker speaking ("speaker's face in a video sequence", NEFIAN, paragraph [0014]), wherein the neural network is trained to detect and monitor a face of the speaker ("face detection", NEFIAN, paragraph [0014]).

23. Regarding **claim 5**, NEFIAN further teaches filtering the detected face of the speaker to detect movement or lack of movement in a mouth of the speaker ("after the face is detected, mouth region discrimination is usual", NEFIAN, paragraph [0015]).

24. Regarding **claim 6**, NEFIAN further teaches that matching further includes comparing portions of the captured visual features against portions of the captured audio during a same time slice ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]).

25. Regarding **claim 9**, KATSUMI teaches all of the limitations of claim 8.

However, KATSUMI does not disclose modeling the analysis for subsequent interactions with the first and second speakers.

In the same field of audiovisual processing, NEFIAN teaches modeling the analysis for subsequent interactions with the first and second speakers ("the result is a model for the underlying process", NEFIAN, paragraph [0033]).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual modeling method of NEFIAN with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

26. Regarding **claim 10**, NEFIAN further teaches that analyzing further includes processing a neural network ("neural network", NEFIAN, paragraph [0014]) for detecting faces of the first and second speakers ("speaker's face in a video sequence", NEFIAN, paragraph [0014]) and processing vector classifying algorithms to detect when the first and second speakers' respective mouths are moving or not moving (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations is performed on the mouth regions).

27. Regarding **claim 14**, NEFIAN further teaches that matching further includes identifying time dependencies associated with when selective portions of the electronic video were monitored and when selective portions of the audio were captured ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]).

28. Regarding **claim 17**, KATSUMI teaches all of the limitations of claim 15.

However, KATSUMI does not disclose that the instructions interact with a neural network to detect a face of the speaker from the captured video.

In the same field of audiovisual processing, NEFIAN teaches instructions that interact with a neural network ("neural network", NEFIAN, paragraph [0014]) to detect a face of the speaker from the captured video ("speaker's face in a video sequence", NEFIAN, paragraph [0014]).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual matching method of NEFIAN with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

29. Regarding **claim 18**, NEFIAN further teaches that the instructions interact with a pixel vector algorithm to detect when a mouth associated with the face moves or does not move within the captured video (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions).

30. Regarding **claim 19**, NEFIAN further teaches that the instructions generate parameter data ("audio processing and visual feature extraction", NEFIAN, paragraph [0012]) that configures a Bayesian network ("video data must be fused with audio data using... a coupled hidden Markov model [HMM]", NEFIAN, paragraph [0023], where the HMM is a dynamic Bayesian network) which models subsequent interactions with the

speaker ("the result is a model for the underlying process", NEFIAN, paragraph [0033]) to determine when the speaker is speaking and to determine appropriate audio to associate with the speaker speaking in the subsequent interactions ("speech recognition", NEFIAN, paragraph [0023]).

31. Regarding **claim 23**, KATSUMI teaches all of the limitations of claim 20.

However, KATSUMI does not disclose a neural network or vector matching algorithm.

In the same field of audiovisual processing, NEFIAN teaches instructions for:

processing a neural network ("neural network", NEFIAN, paragraph [0014]) to detect a face of the speaker ("speaker's face in a video sequence", NEFIAN, paragraph [0014]); and

processing a vector matching algorithm to detect movements of the mouth of the speaker within the detected face (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual matching method of NEFIAN with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

32. Regarding **claim 26**, KATSUMI teaches all of the limitations of claim 25.

However, KATSUMI does not disclose that the apparatus is used to configure a Bayesian network which models the speaker speaking.

In the same field of audiovisual processing, NEFIAN teaches that the apparatus is used to configure a Bayesian network which models the speaker speaking ("video data must be fused with audio data using... a coupled hidden Markov model [HMM]", NEFIAN, paragraph [0023], where the HMM is a dynamic Bayesian network).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual modeling method of NEFIAN with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

33. Regarding **claim 27**, NEFIAN further teaches that the face detection logic comprises a neural network ("neural network", NEFIAN, paragraph [0014]).

34. **Claim 12** is rejected under 35 U.S.C. 103(a) as being unpatentable over Katsumi Patent No.: US 6,369,846 ("KATSUMI") in view of Van Schyndel Patent No.: US 5,940,118 ("VAN SCHYNDEL").

35. Regarding **claim 12**, KATSUMI teaches all of the limitations of claim 8.

However KATSUMI does not specifically disclose suspending the capturing of audio when the analysis does not detect the mouths moving for the first and second speakers.

In the same field of audiovisual processing, VAN SCHYNDEL teaches suspending the capturing of audio when the analysis does not detect the mouths moving for the first and second speakers ("uses optical information to optimally select and/or steer a microphone array in the direction of the talker", VAN SCHYNDEL, column 2, lines 55-58, meaning audio is not captured for someone who is not speaking).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use selectable microphones of VAN SCHYNDEL with the speaker determination system of KATSUMI in order to not restrict a talker's movement or position (VAN SCHYNDEL, column 2, lines 60-61).

36. **Claim 7** is rejected under 35 U.S.C. 103(a) as being unpatentable over Katsumi Patent No.: US 6,369,846 ("KATSUMI") in view of Nefian Pub. No.: US 2003/0212557 ("NEFIAN") and further in view of Van Schyndel Patent No.: US 5,940,118 ("VAN SCHYNDEL").

37. Regarding **claim 7**, the combination of NEFIAN and KATSUMI teach all the limitations of claim 1.

However, NEFIAN and KATSUMI do not specifically disclose suspending the capturing of audio during periods where select ones of the captured visual features indicate that the speaker is not speaking.

In the same field of audiovisual processing, VAN SCHYNDEL teaches suspending the capturing of audio during periods where select ones of the captured

visual features indicate that the speaker is not speaking ("uses optical information to optimally select and/or steer a microphone array in the direction of the talker", VAN SCHYNDEL, column 2, lines 55-58, meaning audio is not captured for someone who is not speaking).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use selectable microphones of VAN SCHYNDEL with the speaker determination system of KATSUMI and audiovisual matching method of NEFIAN in order to not restrict a talker's movement or position (VAN SCHYNDEL, column 2, lines 60-61).

Conclusion

38. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. A list of the pertinent prior can be found on the included form PTO-892 List of References Cited.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Joel Stoffregen whose telephone number is (571) 270-1454. The examiner can normally be reached on Monday - Friday, 9:00 a.m. - 6:30 p.m..

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Patrick Edouard can be reached on (571) 272-7603. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

JS

PATRICK N. EDOUARD
SUPERVISORY PATENT EXAMINER